The Generative AI Paradox: "What It Can Create, It May Not Understand"

Year: 2023 | Citations: 102 | Authors: Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li

Abstract

The recent wave of generative AI has sparked unprecedented global attention, with both excitement and concern over potentially superhuman levels of artificial intelligence: models now take only seconds to produce outputs that would challenge or exceed the capabilities even of expert humans. At the same time, models still show basic errors in understanding that would not be expected even in non-expert humans. This presents us with an apparent paradox: how do we reconcile seemingly superhuman capabilities with the persistence of errors that few humans would make? In this work, we posit that this tension reflects a divergence in the configuration of intelligence in today's generative models relative to intelligence in humans. Specifically, we propose and test the Generative AI Paradox hypothesis: generative models, having been trained directly to reproduce expert-like outputs, acquire generative capabilities that are not contingent upon -- and can therefore exceed -- their ability to understand those same types of outputs. This contrasts with humans, for whom basic understanding almost always precedes the ability to generate expert-level outputs. We test this hypothesis through controlled experiments analyzing generation vs. understanding in generative models, across both language and image modalities. Our results show that although models can outperform humans in generation, they consistently fall short of human capabilities in measures of understanding, as well as weaker correlation between generation and understanding performance, and more brittleness to adversarial inputs. Our findings support the hypothesis that models' generative capability may not be contingent upon understanding capability, and call for caution in interpreting artificial intelligence by analogy to human intelligence.